NIST LoReHLT 2016 Evaluation Plan

Last Updated March 30, 2016

1 Introduction

The DARPA Low Resource Languages for Emergent Incidents (LORELEI) Program seeks to develop human language technology (HLT) that can support rapid and effective response to emerging incidents where the language resources are very limited. As such, LORELEI aims to develop capabilities that can extract knowledge from foreign language sources quickly. This document describes the evaluation specifications of the component evaluation conducted by NIST to assess the performance and track the progress made.

Participation in the NIST Low Resource Human Language Technology (LoReHLT) evaluation is required for all DARPA LORELEI performers responsible for the relevant component technologies in LORELEI. The evaluation is also open to all researchers who find the evaluation tasks of interest. There is no cost to participate. However, participants are expected to attend a post-evaluation workshop to present and discuss their systems and results at their own expense. Information and updates about the component evaluation will be posted to the NIST LoReHLT website[1].

2 Evaluation Tasks

There are three evaluation tasks. LORELEI performers are required to participate in the tasks as outlined by their Statement of Work. Volunteer participants can participate in any and all tasks.

- **Machine Translation (MT)** – for each document, automatically translate it from a given incident language (IL) to English. For MT specific requirements, see Section 12.
- **Situation Frame (SF)** – the SF task specifics are still being discussed, but will be defined in Section 13.
- **Named Entity Recognition (NER)**[2] – for each document, identify and classify named mentions of PER, GPE, ORG, LOC entities. For NER specific requirements, see Section 14.

3 Training Conditions

For each evaluation task, there are two training conditions (constrained and unconstrained) that differentiate the amount/source of incident language-related training material without preventing/excluding multilingual resources and technology. The intent of the *constrained* training condition is to test multilingual systems that are re-targeted to an incident language using a fixed amount of incident language materials. Teams should consult with NIST if their approach is not easily classifiable.

[1] http://www.nist.gov/itl/iad/mig/lorehlt16.cfm

[2] This task is for year 1 only. In subsequent years (2+), the task will be Entity Discovery and Linking (EDL)

- **Constrained** – The constrained data condition limits the incident language material used to train/adapt the tested technology to only those distributed according to Section 5 (IL Data) and Section 6 (Native Language Informants). No other incident language materials, i.e., parallel text, speech corpora, etc. are permitted but knowledge gained from the Native Language Informant is permitted. Prior to the evaluation period, which begins with the announcement of the IL, teams can assemble multilingual resources/technologies/etc. to use during the evaluation so long as they are multilingual-focused in nature. Serendipitous included incident language data in a multilingual system is allowed and must be documented in the system description. The use of mono- and bi-lingual resources is allowed so long as they do not include the incident language. The Constrained training condition is **required for each task participated**.
- **Unconstrained** – The unconstrained condition removes the limitations of the constrained condition. Teams can use additional, publicly available, incident language materials obtained before or after the IL announcement from an epoch before or after the incident. Teams can use pre-existing, mono-lingual technologies for the incident language. Teams can use additional Native Language Informant time beyond the limits in Section 6. The teams must document the additional data and technologies in their system description. The unconstrained training condition **optional but encouraged**.

4 Baseline Training Data

For each evaluation task, a set of non-IL data resources will be provided by the LDC for training prior to the evaluation period. To obtain this data, volunteer participants must register to participate and sign the license agreement which can be found on the NIST LoReHLT website.

5 Evaluation Data

5.1 Component Definition & Release Plan

All three evaluation tasks have the same data components and release plan. The LDC releases the Incident Language (IL) data and English Scenario Model in an encrypted format (see 5.3), and NIST releases the appropriate decryption key(s) at the appropriate stages. Participants must complete all three checkpoints for their submissions to be considered complete. The stages are:

- Pre-IL Announcement (before the IL Announcement)
  - **Set 0**: Encrypted pre-incident IL training data released
  - **Set 1**: Encrypted incident/post-incident IL training data set 1 released
  - **Set 2**: Encrypted incident/post-incident IL training data set 2 released
  - **Set S**: Encrypted incident/post-incident English Scenario Model released
  - **Set E**: Encrypted incident/post-incident IL evaluation data released
- IL Announcement
  - Identity of IL announced
  - Decryption keys for **set 0** and **set E** released
- Evaluation Checkpoint 1
  - Train with data from **set 0** begins at IL Announcement
  - Evaluation Checkpoint 1 submission due 7 days after IL Announcement

    - Decryption key for **set 1** released 7 days after IL Announcement and after submission to Evaluation Checkpoint 1 made
- Evaluation Checkpoint 2
    - Train with data from **set 0** begins at IL Announcement
    - Train with data from **set 1** begins after the Evaluation Checkpoint 1 submission deadline and the team makes a submission
    - Evaluation Checkpoint 2 submission due 14 days after IL Announcement
    - Decryption key for **set 2** released 14 days after IL Announcement and after submission to Evaluation Checkpoint 2 made
- Evaluation Checkpoint 3
    - Train with data from **set 0** begins at IL Announcement
    - Train with data from **set 1** begins after the Evaluation Checkpoint 1 submission deadline and the team makes a submission
    - Train with data from **set 2** and **set S** begins after the Evaluation Checkpoint 2 submission deadline and the team makes a submission
    - Evaluation Checkpoint 3 submission due 30 days after IL Announcement

5.2 Data Description

The composition of the five datasets (**set 0, set 1, set 2, set S, and set E**) are listed in the table below. The given target data volume is **approximate** and depends on data availability. If the amount for a genre is short of the target, LDC will substitute with another genre. "Kw" refers to multiples of 1000 words.

Set 0 – pre-incident epoch

Category I Resources[3]
• Monolingual Source Text:
  ○ ~100Kw newswire
  ○ ~75Kw discussion forum/blog
  ○ ~50Kw Twitter/SMS
• Parallel Text[4]:
  ○ ~100Kw newswire
  ○ ~100Kw discussion forum/blog
  ○ ~100Kw Twitter/SMS
• Parallel Dictionary (~10,000 stems/lemmas)
 	
Category II Resources (any 5 of the following):
• parallel dictionary IL --> non-English
• monolingual IL dictionary
• monolingual IL grammar book
• parallel English --> IL grammar book
• monolingual IL primer book
• monolingual IL gazetteer
• parallel IL --> English gazetteer

Set 1 – incident/post-incident epoch
Monolingual Source Text – 1/3 of leftover after **set E** is met
Set 2 – incident/post-incident epoch
Monolingual Source Text – 2/3 of leftover after **set E** is met
Set S – incident/post-incident epoch
English Scenario Model – approximately 50Kw, genre balance will vary based on availability
Set E – incident/post-incident epoch
Source Text: • ~100Kw newswire • ~50Kw discussion forum/blog • ~50Kw Twitter/SMS

5.3 Data Encryption

The dataset described above will be encrypted using OpenSSL. NIST has created a package with instructions on how to encrypt and decrypt the data using some sample data. The package can be downloaded from the LoReHLT website.

[3] One of the category I resources (monolingual text, parallel text, or parallel dictionary) must exceed the minimum target by 500%.

[4] The parallel text is found/harvested data and automatically aligned, not created (e.g. via professional translation agency or crowdsourcing). ~300Kw comparable may be substituted for every 100Kw parallel if parallel text is not available.

6 Native Informant Resources

During the evaluation period, participants are allowed the use of a native informant (NI) in their system development. The LORELEI performers will be provided the native informant by their sponsor[5] through the data provider Appen. The native informant will be available remotely via telephone or internet connection. Volunteer participants, if they wish to use a native informant, have to supply their own at their own cost and are free to determine how they communicate with their informant. However, consultation with the informant, by LORELEI performers and volunteer participants, must abide by the following guidelines:

- Informant can be a native speaker of the IL but cannot be a professional linguist.
- It is up to the individual teams to determine how they will make use of the informant. However, **the evaluation data must remain unseen and sequestered, and all probings of the evaluation data are prohibited**. The teams must document how they have used the informant (e.g. producing additional resources for training, etc.).
- If a member(s) of the developer's team also happens to be a native speaker of the IL, this information must also be documented.
- For the constrained training condition, consultation with the informant is limited to the number of hours listed below for each task a team participates regardless of how many submissions. If the use of the native informant exceeds the number of hours given, the submissions are considered to be in the unconstrained training track.
    - 1 hour for Evaluation Checkpoint 1
    - 5 hours for Evaluation Checkpoint 2 (4 hours if 1 hour was used in Checkpoint 1)

7 Evaluation Protocol

7.1 Evaluation Account

All participants are required to sign up for an evaluation account on the NIST LoReHLT evaluation web site as all evaluation activities will be conducted via the evaluation account. Go to https://lorehlt.nist.gov to sign up for an account. Participants will need a valid email address and choose a password that is at least 12 characters long including uppercase and lowercase letters, numbers, and special characters.

After signing up and confirming the account, participants will be asked to associate themselves to a site[6] (or create their site if it does not exist). The first person that creates the site will be deemed the site representative and will be asked to associate his site to a team[7] (or create their team if it does not exist). The first person that creates the team will be deemed the team representative. The team representative will be asked to register for any or all of the evaluation tasks. After the task registration, if the site is not

[5] LORELEI performers will be provided NI time by their sponsor only for the amount given above. If teams want additional time, they must make their own arrangement at their own cost.
[6] A site is defined to be a single organization participating in the evaluation.
[7] A team is defined to be a group of organizations collaborating on a task in the evaluation.

a LORELEI performer, the site representative will be asked to sign the LDC license. The LDC will confirm the license or LORELEI performer status[8] and release the appropriate data to the site.

7.2 Submission Requirements

Each team is required to participate in the constrained training condition and is encouraged to participate in the unconstrained training condition. One of the goals of the LoReHLT evaluation is to track system performance over time. As such teams are required to submit at least one ensemble per the training condition participated. An ensemble is defined to be a set of three submissions, one at each checkpoint. Each team can submit up to five ensembles for each training condition and must designate one as the primary ensemble for cross-team comparisons. At each checkpoint, teams are required to provide a short description of their submissions after they upload their system output. At the conclusion of the evaluation, teams are required to submit a more formal system description[9]. The final results will be released to teams who submit a system description. The system descriptions will be compiled into the workshop proceedings. Team can download the template for the system description on the LoReHLT website.

At submission time, teams will be asked to associate their submissions with the appropriate category.

- Task: MT, SF, or NER
- Training condition: Constrained or Unconstrained
- System type: one of primary ensemble, contrast1 ensemble, …, or contrast4 ensemble

For the specific task system output format, refer to the task specific requirements.

8 Evaluation Rules and Requirements

The evaluation is an open evaluation where the test data is sent to the participants who will process and submit the output to NIST. As such, the participants have agreed to process the data in accordance with the following rules:

- The participant agrees not to investigate the evaluation data. Both human/manual and automatic probing of the evaluation data is prohibited to ensure that all participating systems have the same amount of information on the evaluation data.
- The participant agrees to abide by the terms guiding the use of the native informant[10].
- The participant agrees to process at least the constrained training track for each of the selected tasks.
- The participant agrees to complete all three checkpoints to be considered a complete submission for each selected task and training track combination.

[8] LORELEI performers already signed a separate license with the LDC and do not have to sign the LoReHLT data license. However, LDC will need to confirm that the identity is valid.
[9] the template for the system description can be found on the LoReHLT website http://www.nist.gov/itl/iad/mig/lorehlt16.cfm
[10] contact NIST at lorehlt_poc@nist.gov if this presents a problem

- The participant agrees to participate in the dry run exercise to ensure evaluation readiness.
- The participant agrees to attend a post-evaluation workshop to present and discuss his/her systems. Failure to attend the workshop may result in participant being denied from participating in future evaluations.
- The participant agrees to the rules governing the publication of the results.

9 Guidelines for Publication of Results

This evaluation follows an open model to promote interchange with the outside world. At the conclusion of the evaluation cycle, NIST will create a report that documents the evaluation. The report will be posted on the NIST web space and will identify the participants and the scores from various metrics achieved for task.

The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

9.1 Rules Governing Publication of Evaluation Results

The rules governing the publication of the LoReHLT evaluation results are similar to those used in other MIG evaluations.

- Participants are free to publish results for their own system, but participants must not publicly compare their results with other participants (ranking, score differences, etc.) without explicit written consent from the other participants.
- While participants may report their own results, participants may not make advertising claims about winning the evaluation or claim NIST endorsement of their system(s). Per U.S. Code of Federal Regulations (15 C.F.R. § 200.113): *NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.*
- All publications must contain the following NIST disclaimer:

  NIST serves to coordinate the evaluations in order to support research and to help advance the state- of-the-art. NIST evaluations are not viewed as a competition, as such results reported by NIST are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U.S. Government.

10 Dry Run

All participants are required to participate in a dry run evaluation to demonstration evaluation readiness. The purpose of the dry run is to iron out all of the bugs in the evaluation pipeline, not to gauge any

system's capability. The dry run will follow the exact protocol of the official evaluation except that LORELEI performers will have access to the native informant only for checkpoint 1 for 1 hour[11].

11 Schedule

Milestone	***Date***
Evaluation plan published	Dec 11, 2015
Registration opens	Feb 19
6-month PI meeting in San Antonio, TX (LORELEI performers only)	Feb 24 – 26, 2016
Registration deadline (Volunteer participants)	May 2
Required dry run begins	Jun 2
Required dry run ends	Jun 8
Teams receive encrypted IL data	Jun 29
IL Announcement Decryption keys for **set 0** and **set E** distributed	Noon Jul 6
Evaluation Checkpoint 1 submissions due Decryption key for **set 1** distributed after submission made	Noon Jul 13
Evaluation Checkpoint 2 submission due Decryption key for **set 2** distributed after submission made	Noon Jul 20
Evaluation Checkpoint 3 submission due	Noon Aug 3
System description due	Aug 9
Preliminary results released if system description is received	Aug 10
Human Assessments	TBD
Human Assessment results released	TBD
2-day NIST evaluation workshop co-locate with DARPA PI meeting (Volunteer participants; optional for LORELEI performers)	Aug 28 – 29
3-day DARPA PI meeting (LORELEI performers only)	Aug 30 – Sep 1
Meeting to discuss Human Assessment Results	~Nov

[11] Pending contract award prior to the dry run

12 Machine Translation (MT) Evaluation Specifications

12.1 Task Definition

Given a text document in the incident language, the MT system is required to automatically translate the document's content into English. The entire test set must be translated, even though only a subset of it will be scored in the machine translation evaluation.

12.2 Performance Measurements

BLEU and METEOR will be the primary metrics in Phase 1. BLEU and METEOR scores will be calculated at each checkpoint. Scoring will be done against four human reference translations. Scoring will be done preserving case. Other normalizations may be implemented for scoring purposes as necessary for the domains and data encountered.

NIST will investigate additional automatic approaches geared towards measurement of successful translation of content.

12.3 Input and Output Format

LDC will provide the source data in LTF format conforming to the LTF DTD referenced inside the test files. The MT system output must be well-formed and validate against the following DTD which is based on simplified formatting requirements of prior NIST OpenMT evaluations: ftp://jaguar.ncsl.nist.gov/mt/resources/lorehlt-mt-v1.2.dtd

Below are samples of the source and system translation files. The text of the segments has been replaced by English placeholders.

Sample source file, LTF format:

```
<?xml version="1.0"?>
<!DOCTYPE LCTL_TEXT SYSTEM "ltf.v1.5.dtd">
<LCTL_TEXT lang="URD" source_file="NW_JAN_URD_023735_20060331.rsd.txt"
           source_type="web_news" author="LDC" encoding="UTF-8">
  <DOC id="NW_JAN_URD_023735_20060331" lang="URD">
    <TEXT>
      <SEG id="NW_JAN_URD_023735_20060331-1" start_char="3"
           end_char="17">
        <ORIGINAL_TEXT>source segment1</ORIGINAL_TEXT>
        <TOKEN id="NW_JAN_URD_023735_20060331-1-1"
               start_char="3" end_char="8">source</TOKEN>
        <TOKEN id="NW_JAN_URD_023735_20060331-1-2" start_char="10"
               end_char="17">segment1</TOKEN>
      </SEG>
    <SEG id="NW_JAN_URD_023735_20060331-2" start_char="19"
         end_char="33">
      <ORIGINAL_TEXT>source segment1</ORIGINAL_TEXT>
      <TOKEN id="NW_JAN_URD_023735_20060331-2-1" start_char="19"
             end_char="24">source</TOKEN>
      <TOKEN id="NW_JAN_URD_023735_20060331-2-2" start_char="26"
             end_char="33">segment2</TOKEN>
    </SEG>
```

```
…
    </TEXT>
  </DOC>
</LCTL_TEXT>
```

Sample MT system translation file:

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE mteval SYSTEM "lorehlt-mt-v1.2.dtd">
<mteval>
  <tstset>
    <doc docid="NW_JAN_URD_023735_20060331">
      <seg id="NW_JAN_URD_023735_20060331-1">system segment1</seg>
      <seg id="NW_JAN_URD_023735_20060331-2">system segment2</seg>
      …
    </doc>
  </tstset>
</mteval>
```

13 Situation Frame (SF) Evaluation Specifications

As the details of the Situation Frame task are still being discussed, the content of this section is forthcoming.

14 Named Entity Recognition (NER) Evaluation Specifications

14.1 Task Definition

Given a document in the incident language, an NER system is required to automatically identify and classify entity mentions into pre-defined entity types. Note only named mentions are targeted. The entity types in LORELEI/LORE tasks are listed as follows: (To be aligned with LDC NER annotations, we are planning to follow their definitions, so the following definitions may subject to change. A pointer to LDC's annotation guidelines will be given later.)

- Person (PER): Person entities are limited to humans identified by name, nickname or alias.
- Geo-political Entity (GPE): GPE entities are composite entities, such as cities, provinces, countries, meaning there are several criteria that must be present to make something a GPE. GPEs consist of (1) a physical location, (2) a government, and (3) a population. All three of these elements must be present for an entity to be tagged as a GPE, as in: United States, China, Pennsylvania, Philadelphia
- Organization (ORG): Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure.
- Location (LOC): Location entities include geographically defined places such as bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments (namely, facilities defined in previous programs such as ACE and KBP EDL).

Other types of named entities like events, animals, inanimate objects and monetary units will not be annotated.

14.2 Performance Measurements

Scoring metrics from TAC KBP2014/2015 tasks will be extended to the NER tasks. System output will be computed against the gold annotation output for precision (P), recall (R) and their balanced harmonic mean (F1). The official metric will be based on exact mention boundary matches, that is, a name mention is correctly labeled if its entity type and start/end offsets match those of a reference name mention. Specifically, we report these three metrics (P, R and F1) for strong_typed_mention from TAC2015 EDL measurements. The detailed description of TAC EDL scoring metrics is in section 2.2 in the overview paper: http://nlp.cs.rpi.edu/paper/kbp2015.pdf.

In addition to the exact match metric, we award systems for partial matches according to the degree of character overlap between system and key names for diagnostic analysis. The partial match scoring algorithm has two parameters: the recall overlap strategy and the precision overlap strategy.

- The per-name recall score of a name in the answer key is the fraction of its characters which overlap with the system name set according to the recall overlap strategy parameter. For the "MAX" strategy, this will be the characters overlapping with the single system name with maximum overlap. For the "SUM" strategy, this will be the number of its characters which overlap with any system mention.
- The recall score for a system is the mean of the per-name recall scores for all names in the answer key.
- The per-name precision score of a name in the answer key is the fraction of its characters overlapped by the reference set, where ”overlapping" is determined by the precision overlap strategy in the same manner as above for recall.
- The precision score for a system is the mean of the per-name precision scores for all names in the answer key.

We will report scores for all four parameter combinations. The scorer is available at: https://github.com/wikilinks/neleval.

14.3 Input and Output Format

The input and output formatting requirements for NER evaluation will be very similar to the most recent TAC2015 EDL evaluation. For more details, please refer to TAC2015 EDL task guidelines Section 2.3: http://nlp.cs.rpi.edu/kbp/2015/kbp2015edl_taskspec_v1.1.pdf

14.3.1 Input

The input for NER is a set of Incident Language documents provided by LDC. The input file is a sentence segmented and tokenized *.ltf file, with an example as follows:

```
<?xml version="1.0" encoding="UTF-8"?>

<!DOCTYPE LCTL_TEXT SYSTEM "ltf.v1.5.dtd">

<LCTL_TEXT>
```

<DOC id="DF_AOA_TUR_0000116_20140900" tokenization="tokenization_parameters.v2.0" grammar="none" raw_text_char_length="1868" raw_text_md5="a74c236c46075766951103ddb5ead01b">

<TEXT>

<SEG id="segment-0" start_char="0" end_char="36">

<ORIGINAL_TEXT>Merhabalar, Oğlum Eylül 2008 doğumlu.</ORIGINAL_TEXT>

<TOKEN id="token-0-0" pos="word" morph="none" start_char="0" end_char="9">Merhabalar</TOKEN>

<TOKEN id="token-0-1" pos="punct" morph="none" start_char="10" end_char="10">,</TOKEN>

<TOKEN id="token-0-2" pos="word" morph="none" start_char="12" end_char="16">Oğlum</TOKEN>

<TOKEN id="token-0-3" pos="word" morph="none" start_char="18" end_char="22">Eylül</TOKEN>

<TOKEN id="token-0-4" pos="number" morph="none" start_char="24" end_char="27">2008</TOKEN>

<TOKEN id="token-0-5" pos="word" morph="none" start_char="29" end_char="35">doğumlu</TOKEN>

<TOKEN id="token-0-6" pos="punct" morph="none" start_char="36" end_char="36">.</TOKEN>

</SEG>

</TEXT>

</DOC>

</LCTL_TEXT>

For offset counting, the beginning character of the first token has index 0 for the document. The start offset must be the index of the first character in the corresponding string, and end offset must be the index of the last character of the string (therefore, the length of the corresponding mention string is endoffset – startoffset + 1).

14.3.2 Output

An NER system is required to automatically generate an output file, which contains one line for each mention, where each line has the following tab-delimited fields. Please note that while the format is identical to that of TAC2014/2015 EDL, some fields will just be place holders as noted below. Using the same format eliminates needs for making changes to the scorer code. Besides, full EDL is expected in year 2 and beyond.

Field 1: system run ID, unique team_id to identify each team and their runs

Field 2: mention ID, unique for each entity name mention

Field 3: mention head string, the full head string of the entity mention

Field 4: document ID: mention head start offset – mention head end offset, an ID for a document in the source corpus from which the mention head was extracted, the starting offset of the mention head, and the ending offset of the mention head.

Field 5: NIL (in future this is a place holder for reference KB link entity ID)

Field 6: entity type: {GPE, ORG, PER, LOC} type indicator for the entity

Field 7: all should be of type {NAM}

Field 8: a confidence value. Set to 1 for Phase I. In the future this field will be replaced by entity linking confidence score.